Operations Research for Health Care 5 (2015) 21–27

Contents lists available at ScienceDirect

Operations Research for Health Care

journal homepage: www.elsevier.com/locate/orhc

Modeling the effect of short stay units on patient admissions

Maartje E. Zonderland [a,*], Richard J. Boucherie [a], Michael W. Carter [b], David A. Stanford [c]

[a] Stochastic Operations Research & Center for Healthcare Operations Improvement and Research, University of Twente, Postbox 217, 7500 AE Enschede, The Netherlands

[b] Centre for Research in Healthcare Engineering, University of Toronto, 5 King's College Road, Toronto ON, M5S 3G8, Canada

[c] Department of Statistical and Actuarial Sciences, University of Western Ontario, 1151 Richmond Street North, London ON, N6A 5B7, Canada

ARTICLE INFO

Article history:
Received 1 May 2013
Accepted 21 April 2015
Available online 29 April 2015

Keywords:
Capacity planning
Emergency department
Length of stay
Patient admissions
Queuing theory
Short stay unit

ABSTRACT

Two purposes of Short Stay Units (SSU) are the reduction of Emergency Department crowding and increased urgent patient admissions. At an SSU urgent patients are temporarily held until they either can go home or transferred to an inpatient ward. In this paper we present an overflow model to evaluate the effect of employing an SSU on elective and urgent patient admissions.

1. Introduction

Emergency Department (ED) crowding is an increasing problem, resulting in an increased length of stay and prolonged waiting times for patients. Also, ED crowding may result in increased mortality rates and lower quality of care [1]. These problems are not only caused by an aging population [2], a higher demand for acute care [3], and the inability to transfer patients to inpatient beds [3,4], but also by hospital restructuring leading to fewer inpatient beds and more ambulatory care [5].

The main purpose of a Short Stay Unit (SSU) is to temporarily admit suitable ED patients, in order to 'improve the quality of medical care through extended observation and treatment, while reducing inappropriate admissions and health care costs' [6]. Another benefit of an SSU is to act as a temporary holding facility during peak daily demand in the ED. However, the modeling of such transient systems comprising many elements with an analytical model is quite difficult, and hence we have not pursued this aspect in the present study.

An important incentive in the Netherlands for having an SSU is improving ED patient flow and thus decreasing pressure on ED facilities such as rooms, beds and staff. However, the definition and purposes of SSUs vary across hospitals and countries. When the focus is on handling in an effective and prompt manner patients whom it is believed can be diverted from admission to the wards, terms such as "Medical Assessment Unit" are used. The review papers [7–9] provide a comprehensive overview of definitions and concepts for SSU's. Patient types that can be admitted vary, for example sometimes only medical patients are considered [9]. Patients that need intensive care are usually excluded [7,9]. The maximum length of stay (LOS) at the SSU is usually short, with regular repatriations to the inpatient wards. Transfer epochs can be fixed (for example twice a day) or patients can be transferred once a bed becomes available. ED patients who do not require hospitalization but have to wait for test results or require observation for a short period of time can also be admitted. Given the close monitoring, a staffed bed at this location is usually more expensive than a bed at a regular inpatient ward.

The success of an SSU depends on the overlying organizational structures, together with clear agreements upon transfers to regular inpatient wards, a well-defined chain of command, and access to specialist consultations [7,9]. Currently, there is ample debate as to whether operating an SSU reduces ED crowding (see [1,7,9,10] and the references therein). This is not only related to the ambiguity in the terminology and definitions of the SSU used in practice, but could also be caused by a lack of management information. This makes it very hard to measure the effects of opening an SSU on the ED patient flow. Furthermore, there may be a publication bias since it is common to report only positive experiences [9]. This paper aims at filling this gap in literature by

* Corresponding author.
E-mail addresses: m.e.zonderland@utwente.nl (M.E. Zonderland), r.j.boucherie@utwente.nl (R.J. Boucherie), carter@mie.utoronto.ca (M.W. Carter), stanford@stats.uwo.ca (D.A. Stanford).

http://dx.doi.org/10.1016/j.orhc.2015.04.001

providing a quantitative analysis on the effects of having an SSU, under the scenario where a patient's LOS is independent of the type of bed they occupy.

Since ED treatment is expensive compared to inpatient care, it is financially attractive for hospitals to continue the care process at one of their own inpatient wards, instead of transferring the patient to another hospital after ED treatment. The inpatient wards admit elective patients as well, and therefore it is usually a challenge to reserve inpatient beds, given the uncertainty in urgent arrivals. In this view, the SSU serves as extra buffer capacity for patients who require hospitalization.

A number of papers describe improvement of ED patient flow using simulation techniques [11–14]. A few examples use queuing theory [15–18]. Even though the SSU has been the subject of research quite often in the last decade, we were not able to find an analytical evaluation of its effect in terms of inpatient admissions as we present here.

In Section 2 we present a queuing model designed to evaluate the effect of employing an SSU. Methodologically, we build on the Equivalent Random Method as developed by Wilkinson [19] for systems with a single primary cell generating overflow traffic. The method was adapted by Schehrer [20] to represent systems with multiple primary cells. Additional traffic flows have been introduced in [21,22]. We further extend [21] and combine the results with [22], in order to analyze the ED $\rightarrow$ ward $\rightarrow$ SSU system. Using the model, an illustrative example is analyzed in Section 3. Performance measures such as the number of elective and urgent inpatient admissions are calculated.

2. Model

Fig. 1 illustrates the overflow system for the case of $I = 2$ wards. The inpatient wards and SSU can be modeled as an overflow system, where the inpatient wards are the primary stations (i.e., the wards that generate the overflow of urgent patients) and the SSU is the location where urgent patients are routed when the inpatient ward is full. Urgent patients at the SSU have a common exponentially distributed LOS with rate μ_{ssu}. Urgent patient transfers from the SSU to ward i occur with rate γ_i. We assume that the LOS at ward i for these transfer patients is exponential with mean μ_i^{-1}. Patients not requiring hospitalization have a direct ED $\rightarrow$ SSU routing, arrive with Poisson rate λ_0 and have an exponentially distributed LOS with mean μ_0^{-1}. There are I wards, with capacity c_i, $i = 1, \ldots, I$, and related patient types $1, 2, \ldots, I$. We assume that the LOS at ward i is exponentially distributed with mean μ_i^{-1}, so that the LOS for elective and urgent patients at ward i is the same. Urgent patients arrive at ward i with rate λ_{iu}. If all beds at ward i are occupied, the urgent patient is routed to the SSU. If the SSU, which has a capacity of c_0 staffed beds, is fully occupied as well, the patient is blocked and leaves. Elective patients of type i are blocked when ward i is full. Elective patient demand at the wards, which also incorporates patients from the ICU, is modeled with a Poisson process with rate λ_{ie}.

Remarks on the modeling assumptions

1. Although elective arrivals are in fact scheduled, random fluctuations in the number of scheduled arrivals make the Poisson assumption plausible [23]. Of course, scheduled arrivals can be rescheduled and are not lost as our modeling assumption would suggest. However, the cancellation of their elective procedure may well lead some patients to seek future treatment elsewhere. Thus the blocking measure we calculate to measure the increased cancellation rates resulting from the higher bed utilizations in the wards is a relevant performance measure.

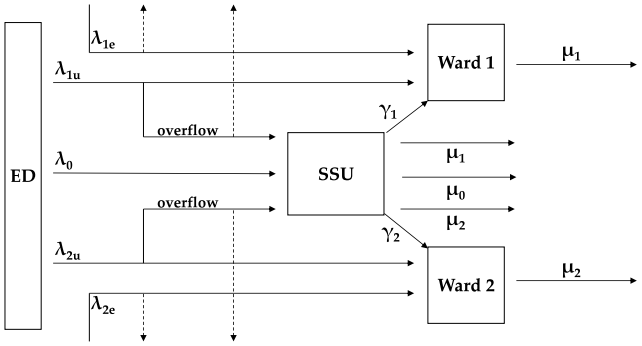

Fig. 1. ED–Ward–SSU patient flow; example with two wards.

2. The rate γ_i is the repatriation rate from the SSU to the wards. If $\gamma_i = 0$, patients of type i are never repatriated from the SSU to ward i. If $\gamma_i = \infty$ patients of type i are immediately repatriated from the SSU once a bed at ward i becomes available.
3. The usage of an SSU could potentially reduce the length of a hospital stay. However, the evidence for this is not yet convincing [6,8]. Therefore the length of stay at the inpatient wards for patients of type i is equal to μ_i^{-1}, regardless of whether the patient was routed via the SSU.
4. Note that it is assumed the patient's route is known once the patient leaves the ED. In practice, sometimes patients are observed in SSU and their final destination is not known at their time of entry. We do, in fact, reflect this phenomenon by treating the time in the SSU for urgent patients as the minimum of the random times until recovery and until they are repatriated to their ward.

2.1. The number of admitted patients

We denote the number of elective and urgent patients present at ward i by n_{ie} and n_{iu} respectively. The number of urgent patients of type i present at the SSU is given by n_{i0}. Patients directly routed to the SSU are considered to be of type 0, and the number of patients of type 0 present at the SSU is denoted with n_{00}. The state space for the overflow system in Fig. 1 is given by

$$S : \Bigg\{ \mathbf{n} = (n_{00}, n_{10}, \ldots, n_{I0}, n_{1e}, \ldots, n_{Ie}, n_{1u}, \ldots, n_{Iu}); \\ n_{ie} + n_{iu} \leq c_i \; \forall i; \sum_{i=0}^{I} n_{i0} \leq c_0; \; n_{ie}, n_{iu}, n_{i0}, n_{00} \geq 0 \; \forall i \Bigg\}. \quad (1)$$

Denote $\pi(\mathbf{n})$ as the equilibrium probability that $\mathbf{n}$ patients are present in the entire ED–Ward–SSU system. To calculate performance measures such as the number of admitted patients at the wards or SSU, the distribution of $\pi(\mathbf{n})$ is required. The latter can be found by solving the following system of global balance equations:

$$\pi(\mathbf{n}) \Bigg[\sum_{i=1}^{I} \lambda_{ie} \mathbb{1}_{n_{ie}+n_{iu}<c_i} + \sum_{i=1}^{I} \lambda_{iu} \left(\mathbb{1}_{n_{ie}+n_{iu}<c_i} + \mathbb{1}_{n_{ie}+n_{iu}=c_i, \sum_{i=0}^{I} n_{i0}<c_0} \right) \\ + \lambda_0 \mathbb{1}_{\sum_{i=0}^{I} n_{i0}<c_0} + \sum_{i=1}^{I} (n_{ie} + n_{iu}) \mu_i + \sum_{i=1}^{I} n_{i0} \mu_i \\ + n_{00} \mu_0 + \sum_{i=1}^{I} n_{i0} \gamma_i \mathbb{1}_{n_{ie}+n_{iu}<c_i} \Bigg] \\ = \Bigg[\sum_{i=1}^{I} \lambda_{ie} \pi(\mathbf{n} - e_{ie}) \mathbb{1}_{n_{ie}>0}$$

$$
\begin{aligned}
&+\sum_{i=1}^{I}\lambda_{iu}\left(\pi(\mathbf{n}-e_{iu})\mathbb{1}_{n_{iu}>0}+\pi(\mathbf{n}-e_{0i})\mathbb{1}_{n_{i0}>0,n_{ie}+n_{iu}=c_i}\right)\\
&+\lambda_0\pi(\mathbf{n}-e_{00})\mathbb{1}_{n_{00}>0}\\
&+\sum_{i=1}^{I}(n_{ie}+1)\mu_i\pi(\mathbf{n}+e_{ie})\mathbb{1}_{n_{ie}+1+n_{iu}\le c_i}\\
&+\sum_{i=1}^{I}(n_{iu}+1)\mu_i\pi(\mathbf{n}+e_{iu})\mathbb{1}_{n_{ie}+n_{iu}+1\le c_i}\Big]\\
&+\sum_{i=1}^{I}(n_{i0}+1)\mu_i\pi(\mathbf{n}+e_{0i})\mathbb{1}_{n_{i0}+1\le c_0}\\
&+(n_{00}+1)\mu_0\pi(\mathbf{n}+e_{00})\mathbb{1}_{n_{00}+1\le c_0}\\
&+\sum_{i=1}^{I}(n_{i0}+1)\gamma_i\pi(\mathbf{n}+e_{0i}-e_{iu})\mathbb{1}_{n_{ie}+n_{iu}+1\le c_i}. \qquad (2)
\end{aligned}
$$

This system of equations can be solved explicitly only for specific values of the system parameters [22]. We modify the analysis from [20] and its generalization presented in [21,22], which required that $\mu_i=\mu_{ssu}$. We adapt it to allow $\mu_i\neq\mu_{ssu}$, and to include the flow with rate λ_{ie} of elective patients of type i. To obtain the mean and variance of the number of patients in the SSU, the model is analyzed for $\gamma_i=0\ \forall i$. Subsequently, for $\gamma_i>0$ the approaches presented in [21,22] are combined, in order to determine the number of patients present at each ward.

2.2. No transfers from the SSU to the wards: $\gamma_i=0$

We first consider the case where SSU patients are not transferred to the wards, i.e., $\gamma_i=0$ (see Fig. 2). Following [20], the mean, $\mathbb{E}_i$, and variance, $\mathbb{V}_i$, of the overflow of urgent patients of type i at the SSU, in case of infinite SSU capacity, can be obtained from the global balance equation (2). Since $c_0=\infty$ and due to the independence of the overflow processes from the wards, $\mathbb{E}_i$ and $\mathbb{V}_i$ can be determined for each ward in isolation. For the blocking probability at the overflow, it does not matter whether a patient residing at ward i is of urgent or elective type. Let $n_i=n_{ie}+n_{iu}$ denote the number of type i patients at ward i, and let $\lambda_i=\lambda_{ie}+\lambda_{iu}$ denote the total arrival rate at ward i. The system of global balance equations simplifies to

$$
\begin{aligned}
&\pi(n_{i0},n_i)\,(\lambda_i+n_i\mu_i+n_{i0}\mu_i)\\
&\quad=\lambda_i\pi(n_{i0},n_i-1)+(n_{i0}+1)\mu_i\pi(n_{i0}+1,n_i)\\
&\qquad+(n_i+1)\mu_i\pi(n_{i0},n_i+1)\quad\text{for } n_i<c_i,\\
&\pi(n_{i0},c_i)\,(\lambda_{iu}+c_i\mu_i+n_{i0}\mu_i)\\
&\quad=\lambda_i\pi(n_{i0},c_i-1)+(n_{i0}+1)\mu_i\pi(n_{i0}+1,c_i)\\
&\qquad+\lambda_{iu}\pi(n_{i0}-1,c_i). \qquad (3)
\end{aligned}
$$

Define the probability generating function of the number of urgent patients of type i present at the SSU, $G_{i,n_i}(z)$, as

$$
G_{i,n_i}(z)=\sum_{n_{i0}=0}^{\infty}\pi(n_{i0},n_i)z^{n_{i0}},\quad |z|<1. \qquad (4)
$$

Multiplication of (3) by $z^{n_{i0}}$ and the summation of the result over $n_{i0}=0,\ldots,\infty$ yields

$$
\begin{aligned}
&[\lambda_i+n_i\mu_i]\,G_{i,n_i}(z)+\mu_i(z-1)\frac{d}{dz}G_{i,n_i}(z)\\
&\quad=\lambda_iG_{i,n_i-1}(z)+(n_i+1)\mu_iG_{i,n_i+1}(z)\quad\text{for } 0\le n_i<c_i,\\
&[\lambda_{iu}(1-z)+c_i\mu_i]\,G_{i,c_i}(z)+\mu_i(z-1)\frac{d}{dz}G_{i,c_i}(z)\\
&\quad=\lambda_iG_{i,c_i-1}(z). \qquad (5)
\end{aligned}
$$

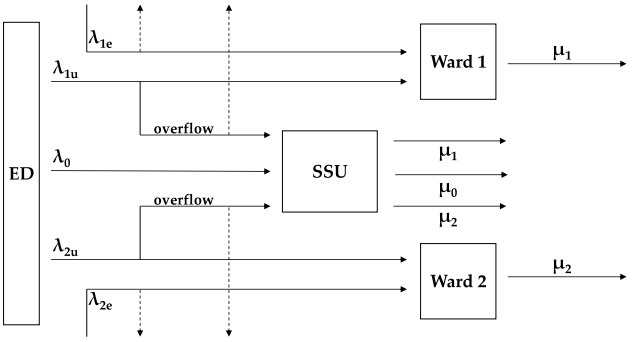

Fig. 2. No transfers to the wards; $\gamma_i=0$. Example with two wards.

Now $\mathbb{E}_i$ and $\mathbb{V}_i$ can be derived from

$$
\begin{aligned}
\mathbb{E}_i&=\sum_{n_i=0}^{c_i}\frac{d}{dz}G_{i,n_i}(z)\Big|_{z=1}\\
\mathbb{V}_i&=\sum_{n_i=0}^{c_i}\frac{d^2}{dz^2}G_{i,n_i}(z)\Big|_{z=1}+\mathbb{E}_i-(\mathbb{E}_i)^2. \qquad (6)
\end{aligned}
$$

Taking the first derivative of (5) and evaluating at $z=1$, gives

$$
\begin{aligned}
&(\lambda_i+n_i\mu_i+\mu_i)\,g_i[n_i]\\
&\quad=\lambda_ig_i[n_i-1]+(n_i+1)\mu_ig_i[n_i+1]\quad\text{for } 0\le n_i<c_i,\\
&(n_i\mu_i+\mu_i)\,g_i[n_i]-\lambda_{iu}\mathbb{P}_i(c_i)=\lambda_ig_i[n_i-1]\quad\text{for } n_i=c_i, \qquad (7)
\end{aligned}
$$

where $g_i[n_i]=\frac{d}{dz}G_{i,n_i}(z)|_{z=1}$ and $\mathbb{P}_i(c_i)$ is the Erlang blocking probability given by

$$
\mathbb{P}_i(c_i)=Erl\left(\frac{\lambda_i}{\mu_i},c_i\right)=\frac{\frac{\left(\frac{\lambda_i}{\mu_i}\right)^{c_i}}{c_i!}}{\sum_{i=0}^{c_i}\frac{\left(\frac{\lambda_i}{\mu_i}\right)^{i}}{i!}}. \qquad (8)
$$

Then $\mathbb{E}_i$ is obtained by the summation of (7) for $n_i=0,\ldots,c_i$:

$$
\mathbb{E}_i=\frac{\lambda_{iu}}{\mu_{ssu}}\mathbb{P}_i(c_i). \qquad (9)
$$

The variance can be calculated accordingly by taking the second derivative of (5) and evaluating at $z=1$:

$$
\begin{aligned}
&(\lambda_i+n_i\mu_i+2\mu_i)\,h_i[n_i]\\
&\quad=\lambda_ih_i[n_i-1]+(n_i+1)\mu_ih_i[n_i+1]\quad\text{for } 0\le n_i<c_i,\\
&(c_i\mu_i+2\mu_i)\,h_i[c_i]-2\lambda_{iu}g_i[c_i]=\lambda_ih_i[c_i-1], \qquad (10)
\end{aligned}
$$

where $h_i[n_i]=\frac{d^2}{dz^2}G_{i,n_i}(z)|_{z=1}$. Summation of the result for $n_i=0,\ldots,c_i$ yields

$$
\mathbb{V}_i=\frac{\lambda_{iu}}{\mu_i}g_i[c_i]+\mathbb{E}_i-(\mathbb{E}_i)^2, \qquad (11)
$$

where $g_i[c_i]$ can be determined recursively from (7) with $g_i[-1]=0$. Note that due to the PASTA property, $\mathbb{P}_i(c_i)$ is the blocking probability of elective patients at ward i.

The direct patient flow arriving at the SSU can be represented by a Poisson distribution with mean $\frac{\lambda_0}{\mu_0}$. The mean $\mathbb{E}_0$ and variance $\mathbb{V}_0$ is therefore given by

$$
\mathbb{E}_0=\mathbb{V}_0=\frac{\lambda_0}{\mu_0}. \qquad (12)
$$

Since the overflow processes from the wards are independent from direct SSU arrivals, the expectation and variation of the aggregated

overflow, $\mathbb{E}$ and $\mathbb{V}$, are given by

$$\mathbb{E} = \sum_{j=0}^{I} \mathbb{E}_j,$$

$$\mathbb{V} = \sum_{j=0}^{I} \mathbb{V}_j. \tag{13}$$

Using the equivalent random method [19], it is now possible to define an equivalent ward with service rate μ_i, which generates overflow traffic to the SSU with the same mean and variance as the i overflow (urgent) and direct streams together. This equivalent ward is constructed solely for the purpose of evaluating the blocking probability of urgent patients. Elective patients are not incorporated in the equivalent ward, since only urgent patients are routed to the SSU. The equivalent ward has load a and capacity C such that [19]

$$aErl(a, C) = \mathbb{E}$$

$$\mathbb{E}\left(1 - \mathbb{E} + \frac{a}{C + 1 + \mathbb{E} - a}\right) = \mathbb{V}. \tag{14}$$

The blocking probability for patients of type i, and for patients of type 0, is [24]

$$K_i = \frac{aErl(a, C + c_0)}{aErl(a, C)} \times \left(v(C, c_0)^{-1} + \frac{\zeta_i - 1}{\zeta - 1}\left(1 - v(C, c_0)^{-1}\right)\right), \tag{15}$$

with $\zeta = \frac{\mathbb{V}}{\mathbb{E}}$, and $\zeta_i = \frac{\mathbb{V}_i}{\mathbb{E}_i}$, which is the peakedness of the separate flows. The variable $v(C, c_0)$ can be determined recursively from

$$v(C, j) = \frac{aj}{aErl(a, C)\,(C + j - a - aErl(a, C)v(C, j-1))},$$
$$j = 1, 2, \ldots$$
$$v(C, 0) = 1. \tag{16}$$

The mean number of urgent patients of type i present at the SSU, $\mathbb{E}[N_{0i}]$, is given by

$$\mathbb{E}[N_{0i}] = \frac{\lambda_{iu}}{\mu_i} Erl\left(\frac{\lambda_i}{\mu_i}, c_i\right)(1 - K_i). \tag{17}$$

The mean number of type 0 patients, $\mathbb{E}[N_{00}]$, equals

$$\mathbb{E}[N_{00}] = \frac{\lambda_0}{\mu_0}(1 - K_0). \tag{18}$$

2.3. Transfers from the SSU to the wards: $0 < \gamma_i < \infty$

We now return to the general case where patient transfers from the SSU back to the wards are allowed. Since we assumed that μ_{ssu} is equal for all patients, γ_i is defined such that $\mu_{ssu} = \mu_i + \gamma_i\ \forall i$. This is a natural choice since one of the purposes of an SSU is provision of short-time treatment. The focus, therefore, lies on either readmission at an inpatient ward (with rate γ_i) or discharge (with rate μ_i). As a result, $\mu_{ssu} \geq \mu_i\ \forall i$, which corresponds exactly to the model given in Fig. 1. The arrival rate at ward i, ν_i, is approximated by the sum of elective and urgent patient arrivals (λ_{ie} and λ_{iu} respectively) and SSU patient transfers [22], $\gamma_i\mathbb{E}[N_{0i}]$

$$\nu_i = \lambda_{ie} + \lambda_{iu} + \gamma_i\mathbb{E}[N_{0i}]. \tag{19}$$

A fraction of this stream, κ_i, overflows to the SSU

$$\kappa_i = \lambda_{iu} + \gamma_i\mathbb{E}[N_{0i}]. \tag{20}$$

Table 1
Parameter values for SSU example.

Parameter	Value	Parameter	Value
c_1	200	μ_0	4
c_2	200	μ_{ssu}	$\frac{2}{3}$
λ_{1e}	26	μ_1	$\frac{1}{5}$
λ_{1u}	14	μ_2	$\frac{1}{4}$
λ_{2e}	37	γ_1	$\frac{7}{15}$
λ_{2u}	13	γ_2	$\frac{5}{12}$

To analyze the model for $\gamma_i > 0$, we replace λ_{iu} by κ_i, and λ_i by ν_i, in Eqs. (7)–(11). We then obtain a system of equations, which can be solved for $\mathbb{E}[N_{0i}]$ using fixed point iteration with initial value $\mathbb{E}[N_{0i}] = 0$ [22].

The mean of the total number of patients present at the SSU, $\mathbb{E}[N_0]$, and the mean of the total number of patients present at ward i, $\mathbb{E}[N_i]$, are given by

$$\mathbb{E}[N_0] = \sum_{i=0}^{I} \mathbb{E}[N_{0i}],$$

$$\mathbb{E}[N_i] = \frac{\nu_i}{\mu_i}\left(1 - Erl\left(\frac{\nu_i}{\mu_i}, c_i\right)\right). \tag{21}$$

The mean occupation, ρ_j, of the SSU ($j = 0$) and the wards ($j \in I$), is given by

$$\rho_j = \frac{E[N_j]}{c_j}. \tag{22}$$

2.4. Immediate transfers from the SSU to the wards: $\gamma_i = \infty$

For completeness, we mention the case where $\gamma_i = \infty$, i.e., patients are immediately transferred from the SSU to the wards. If γ_i were set to infinity, we would obtain a loss network for which an exact expression of the equilibrium distribution could easily be found [22]. This might seem realistic, but in practice transfers do occur after a certain delay due to the limited availability of nursing staff and the patient transportation services.

3. Results

We now use the model from Section 2.3 to analyze a simple example for a hospital with two aggregated wards. Primary ward 1 has a capacity of $c_1 = 200$ beds and admits only medical patients, whose mean LOS is five days (so $\mu_1 = \frac{1}{5}$). The elective patient arrival rate λ_{1e} equals 26 patients per day, and the urgent patient arrival rate λ_{1u} is 14 patients per day, so that the total patient arrival rate $\lambda_1 = 40$. Primary ward 2 admits only surgical patients, with $c_2 = 200$, a mean LOS of four days ($\mu_2 = \frac{1}{4}$), $\lambda_{2e} = 37$, $\lambda_{2u} = 13$, and $\lambda_2 = 50$. Adding capacity by creating so-called overbeds is not allowed.

Patients for observation arrive directly at the SSU with rate $\lambda_0 = 2$ and have a service rate of $\mu_0 = 4$. With this flow we represent patients who require observation for a short period of time (on average six hours in this case). The mean length of stay for urgent patients at the SSU is set to 36 h, so that $\mu_{ssu} = \frac{2}{3}$. Consequently, $\gamma_1 = \mu_{ssu} - \mu_1 = \frac{7}{15}$ and $\gamma_2 = \mu_{ssu} - \mu_2 = \frac{5}{12}$. This implies that patients with a longer LOS should be repatriated more frequently to their ward than patients with a shorter LOS, in order to keep the LOS at the SSU the same for all urgent patients. Table 1 summarizes the parameter values.

Table 2
Results for opening an SSU.

c_0	$\mathbb{P}(B_{1e})$	B_{1e}	$\mathbb{P}(B_{1u})$	B_{1u}	$\mathbb{P}(B_{2e})$	B_{2e}	$\mathbb{P}(B_{2u})$	B_{2u}	EP/y	UP/y
0	0.0544	1.4132	0.0544	0.7609	0.0544	2.0110	0.0544	0.7066	21,745	9319
4	0.0598	1.5547	0.0237	0.3313	0.0580	2.1469	0.0210	0.2733	21,644	9634
6	0.0615	1.5998	0.0145	0.2024	0.0591	2.1885	0.0113	0.1468	21,612	9728
8	0.0629	1.6352	0.0074	0.1035	0.0598	2.2110	0.0061	0.0796	21,591	9788
12	0.0640	1.6652	0.0015	0.0215	0.0603	2.2322	0.0013	0.0169	21,572	9841

Table 3
Results for opening an SSU with load reduced to 95%.

c_0	$\mathbb{P}(B_{1e})$	B_{1e}	$\mathbb{P}(B_{1u})$	B_{1u}	$\mathbb{P}(B_{2e})$	B_{2e}	$\mathbb{P}(B_{2u})$	B_{2u}	EP/y	UP/y
0	0.0280	0.6908	0.0280	0.3720	0.0280	0.9831	0.0280	0.3454	21,234	9100
4	0.0303	0.7491	0.0086	0.1147	0.0296	1.0389	0.0074	0.0909	21,193	9287
6	0.0309	0.7631	0.0042	0.0556	0.0299	1.0496	0.0036	0.0440	21,184	9326
8	0.0312	0.7705	0.0019	0.0251	0.0300	1.0551	0.0016	0.0196	21,179	9346
12	0.0314	0.7756	0.0003	0.0041	0.0301	1.0589	0.0003	0.0032	21,176	9360

Table 4
Results for admitting observation patients.

c_0	$\mathbb{P}(B_{1e})$	B_{1e}	$\mathbb{P}(B_{1u})$	B_{1u}	$\mathbb{P}(B_{2e})$	B_{2e}	$\mathbb{P}(B_{2u})$	B_{2u}	$\mathbb{P}(B_0)$	B_0
4	0.0593	1.5409	0.0265	0.3716	0.0578	2.1377	0.0232	0.3019	21,652	9609
12	0.0640	1.6647	0.0016	0.0228	0.0603	2.2319	0.0014	0.0176	21,573	9840

Table 5
Results for admitting observation patients with load reduced to 95%.

c_0	$\mathbb{P}(B_{1e})$	B_{1e}	$\mathbb{P}(B_{1u})$	B_{1u}	$\mathbb{P}(B_{2e})$	B_{2e}	$\mathbb{P}(B_{2u})$	B_{2u}	$\mathbb{P}(B_0)$	B_0
4	0.0299	0.7375	0.0123	0.1642	0.0294	1.0327	0.0096	0.1183	21,199	9259
12	0.0314	0.7760	0.0001	0.0024	0.0301	1.0592	0.0002	0.0019	21,175	9361

3.1. Opening the SSU

Suppose the hospital considers opening an SSU. We first analyze the situation where only urgent patients are admitted at this ward (so for now, we set $\lambda_0 = 0$). In Table 2 for $c_0 = 0$ (no SSU, i.e., the old situation), and $c_0 = 4, 6, 8, 12$ the blocking probabilities for elective, $\mathbb{P}(B_{ie})$, and urgent patients, $\mathbb{P}(B_{iu})$, are given. The number of rejected elective, B_{ie}, and urgent patients, B_{iu}, is given per ward per day, and the number of admitted elective, EP/y, and admitted urgent, UP/y, patients per year are given.

When $c_0 = 0$, the two wards each act as independent M/G/200/200 loss systems, with blocking probability $Erl(200, 200) = 0.0544$ in both wards. What we see as c_0 increases is that the blocking probability for urgent patients decreases, which was expected since we added capacity for these patients. However, since the hospital is now able to admit more urgent patients, ultimately there is less capacity available at the wards for elective patients which results in suppression of elective demand. An SSU with four beds results in a total of 315 more (9634 vs. 9319) urgent patients admitted per year, but at the same time 101 less elective patients are admitted per year (21,644 vs. 21,745). This might seem minimal, but is equivalent to two canceled patients per week. It follows that the opening of the SSU negatively affects the elective patient flow (once again, we note that the canceled elective patients are not really blocked. However, this blocking measure serves as a useful predictor of the number of patients who might choose to seek future treatment elsewhere as a result of their cancellation).

In order to establish that the phenomenon of increased urgent admissions leading to increased elective cancellations is not an artifact of our chosen load level, we re-ran the configuration at an attenuated load, in which each stream of patients was reduced to 95% of its former value. Thus, the offered load to each ward decreases from 200 patient days per day, to 190. In such large wards, this is enough to cause the blocking levels to drop in a marked fashion; for instance, the Erlang blocking probability essentially is cut in half assuming no SSU beds. Thus, the potential benefit of the SSU beds itself is halved, roughly speaking. As c_0 increases, we still see an increase in the levels of blocking to elective patients; however, the degree is lessened: we see about a 10% increase in elective blocking in Ward 1 when $c_0 = 12$ in Table 3, whereas it had been closer to 20% at full load (from Table 2).

3.2. Admitting observation patients

Following the opening of the SSU, suppose the hospital were to decide that patients from the ED requiring observation should also be admitted here (so that $\lambda_0 > 0$). It is obvious that more beds are required in the SSU to maintain the decreased blocking probabilities for urgent patients (see Table 4 for an example with $\lambda_0 = 2$), but the blocking probabilities for elective patients remain about the same (similar conclusions can be drawn from Table 5 in the case of reduced load).

3.3. Increasing urgent admissions

As mentioned in the Introduction, one of the reasons to open an SSU is to increase the number of urgent patient admissions through the ED. Table 6 shows for various rates of increase, f_u, in the arrival rate of urgent patients, λ_{iu}, the required size of the SSU for which $\mathbb{P}(B_{iu}) \approx 1\%$. Note that $\lambda_0 = 0$.

We see that an increase in the number of urgent patient admissions has a profound effect on the number of elective patient admissions. For example in the case of a 10% increase, the number of elective patient admissions decreases by 3% from 21,745 to 21,065 per year.

3.4. Maintaining the number of elective patient admissions

If the hospital wishes to maintain the number of elective patient admissions, it has two options: increase the number of beds at

Table 6
Results for increasing urgent admissions.

f_u	c_0	$\mathbb{P}(B_{1e})$	B_{1e}	$\mathbb{P}(B_{1u})$	B_{1u}	$\mathbb{P}(B_{2e})$	B_{2e}	$\mathbb{P}(B_{2u})$	B_{2u}	EP/y	UP/y
5%	8	0.0758	1.9716	0.0120	0.1758	0.0688	2.5465	0.0086	0.1176	21,346	10,241
10%	10	0.0910	2.3664	0.0104	0.1601	0.0789	2.9208	0.0067	0.0954	21,065	10,747
20%	12	0.1221	3.1753	0.0126	0.2112	0.0993	3.6731	0.0096	0.1502	20,495	11,694
50%	22	0.2224	5.7827	0.0134	0.2816	0.1660	6.1434	0.0100	0.1955	18,642	14,608

Table 7
Results for maintaining the number of elective patient admissions.

f_u	c_0	c_1	c_2	Added beds
0%	6	203	202	5
5%	6	207	205	12
10%	6	210	208	18
20%	6	215	215	30
50%	8	238	229	67

the wards, or stop repatriating patients from the SSU back to the wards. The latter option would transform the SSU to a long stay ward for urgent patients, and therefore we only analyze the first option. Table 7 gives for each value of f_u the required number of beds at the wards, c_1 and c_2, compared to the initial situation where $c_1 = c_2 = 200$, such that the elective patient blocking probability is maintained at its initial value of ≈5%. For the purposes of these calculations, we have assumed that the SSU is functioning purely as an overflow unit; i.e., $\lambda_0 = 0$.

Since the number of inpatient beds increases, more urgent patients can be admitted directly at the wards, and thus less SSU capacity is required to keep $\mathbb{P}(B_{iu}) \approx 1\%$.

3.5. Sensitivity of the elective patient arrival process

The calculations performed in the preceding tables have been based upon the assumption that all patient arrival streams are Poisson distributed. In reality, elective procedures are scheduled so that the intended instant of an elective arrival is known in advance; it is not randomly distributed. However, as we have already noted in our first remark on our modeling assumptions, there is variability in the number of elective procedures, so that a deterministic arrival process is likewise not an adequate representation of reality. This variation in number from day to day, week to week, and possibly season to season has been seen as a justification in [23] for the use of a Poisson distribution to model it.

Queues featuring both Poisson and non-Poisson arrival streams are such that the respective customers perceive different levels of congestion at their respective arrival instants. For instance, Wolff [25] established the PASTA principle ("Poisson Arrivals See Time Averages"), which applies both to loss and delay systems. Methods to deal with the waiting time distribution in a delay system, as seen by the non-Poisson stream, were developed in the first-come, first-served case by Ott [26], and extended by Stanford [27] in the non-preemptive priority context where the non-Poisson stream has the lowest priority. As a general rule, the more variable the arrival stream, the greater the blocking probability in a loss system, and the longer the waiting time in a delay system.

We performed a simulation-based sensitivity analysis for the shape of the elective arrival process, to compare the situation with Poisson-distributed elective arrivals with another featuring deterministic elective arrivals. Four lengthy simulation experiments, representing 12 replications of 2000-week-long runs with a 200-week-long warm-up period were performed. These were run at the attenuated arrival rate equal to 95% load in both wards, both for the case of $c_0 = 4$ SSU beds and $c_0 = 12$ SSU beds. The results are presented in Tables 8 and 9.

As both tables show, the simulation suggests that introduction of 8 extra SSU beds (from 4 to 12) all but eliminates blocking for the urgent patient streams. The next observation to note is that the confidence intervals from Table 8 are generally supportive of our results presented in Table 3 using the Equivalent Random Method approximation.

We turn next to what the simulation experiments have to say about our assertion, in the absence of any speed up in recovery rendered by an SSU, that the increased access for urgent patients comes at the expense of increased levels of elective stream blocking. When comparing the rows to see the impact of the 8 extra SSU beds, we see that three of the four 95% confidence intervals for the blocking levels are non-overlapping, while the fourth (for elective blocking in ward 2, assuming Poisson-distributed elective arrival streams) has a slight overlap. This would appear to confirm that the reduced blocking to the urgent streams is leading to increased blocking to the elective streams, consistent with the rationale proposed above.

We also observe that the blocking levels for the deterministic arrival streams in Table 9 are about 60% of those seen in the Poisson arrival case in Table 8. The relevant measure for an actual elective stream can be expected to lie between these extremes, for the reasons noted above.

4. Discussion

As stated in the abstract, two purposes of Short Stay Units (SSU) are the reduction of Emergency Department crowding and increased urgent patient admissions. At an SSU urgent patients are temporarily held until they either can go home or are transferred to an inpatient ward. In this paper we have presented an overflow model to evaluate the effect of an SSU on elective and urgent patient admissions, under the scenario that no reduction in overall treatment time can be achieved. The model generalizes Wilkinson's Equivalent Random Method [19–21], to include an extra arrival stream and different service rates at the primary cells (the wards in our example) and at the overflow cell (the SSU). For analytical tractability, we have assumed the LOS at the wards and SSU to be exponentially distributed.

For a hospital with two wards, we have shown that an SSU results in desirable increases in urgent patient admissions, but this comes at the expense of a decrease in the number of elective patient admissions (the elective cancellations we observe are always fewer in number than the increase in urgent admissions). The elective patients are blocked by the increased occupancy of the wards due to urgent patients repatriated from the SSU. In reality, these patients are not in fact blocked; most would merely be rescheduled while others would seek treatment at another hospital as a result of the negative experience due to the cancellation. We have assumed that the urgent patient flow remained constant over time. Furthermore, the added capacity may attract extra urgent patients, which will in turn result in even less capacity for elective patients. To overcome this effect, as well as the SSU capacity created, additional inpatient beds should be added. This in turn results in a decrease of the number of SSU beds required, which makes the SSU a small ward that may be difficult to staff. From a patient flow perspective, the SSU is thus an interesting phenomenon with numerous logistical challenges. In the example we have incorporated only two, very large inpatient wards. In case of more wards with fewer beds, blocking probabilities are more

Table 8
Simulated confidence intervals for blocking assuming Poisson elective arrival streams—the 95% confidence interval lower and upper end points are denoted by subscripts L and U, respectively.

c_0	$\mathbb{P}(B_{1e})_L$	$\mathbb{P}(B_{1e})_U$	$\mathbb{P}(B_{1u})_L$	$\mathbb{P}(B_{1u})_U$	$\mathbb{P}(B_{2e})_L$	$\mathbb{P}(B_{2e})_U$	$\mathbb{P}(B_{2u})_L$	$\mathbb{P}(B_{2u})_U$
4	0.0308	0.0325	0.0074	0.0079	0.0301	0.0313	0.0059	0.0063
12	0.0331	0.0347	0.0001	0.0002	0.0311	0.0322	0.0001	0.0001

Table 9
Simulated confidence intervals for blocking assuming deterministic elective arrival streams—the 95% confidence interval lower and upper end points are denoted by subscripts L and U, respectively.

c_0	$\mathbb{P}(B_{1e})_L$	$\mathbb{P}(B_{1e})_U$	$\mathbb{P}(B_{1u})_L$	$\mathbb{P}(B_{1u})_U$	$\mathbb{P}(B_{2e})_L$	$\mathbb{P}(B_{2e})_U$	$\mathbb{P}(B_{2u})_L$	$\mathbb{P}(B_{2u})_U$
4	0.0185	0.0198	0.0071	0.0077	0.0166	0.0175	0.0054	0.0059
12	0.0201	0.0210	0.0001	0.0002	0.0175	0.0184	0.0001	0.0001

sensitive to an increase in patient arrivals, and thus the blocking effect will remain, and might even worsen.

The question naturally arises as to what circumstances might warrant an SSU or other specialized front-end facility from a performance standpoint. Clearly, if the facility were capable of rendering sufficiently shorter stays at comparable expense, it would be beneficial. The question remains as to how much of a reduction is needed to produce this result. Specialized units that go by names such as Short Stay Units, Medical Assessment Units, and Clinical Decision Units [7–9] aim to attend to suitable patients in a 24–48 time frame without formal admission to the hospital in question; patients still present at that point are typically admitted. Performance analysis of SSUs and similar units under the assumption of a shorter stay in the unit is an extension of the present paper we hope to explore in future work.

Acknowledgments

A significant part of this work was carried out during working visits of the first author to the University of Toronto and the University of Western Ontario. The fourth author acknowledges the support of NSERC (41187-2009 RGPIN) Operating Grant funds to support that stay.

References

[1] N.R. Hoot, D. Aronsky, Systematic review of emergency department crowding: causes, effects, and solutions, Ann. Emerg. Med. 52 (2008) 126–136.
[2] D.C. Roberts, M.P. McKay, A. Shaffer, Increasing rates of emergency department visits for elderly patients in the United States, 1993 to 2003, Ann. Emerg. Med. 51 (2008) 769–774.
[3] J.C. Moskop, D.P. Sklar, J.M. Geiderman, R.M. Schears, K.J. Bookman, Emergency department crowding, part 1—concept, causes, and moral consequences, Ann. Emerg. Med. 53 (2009) 605–611.
[4] A.J. Forster, I. Stiell, G. Wells, A.J. Lee, C. van Walraven, The effect of hospital occupancy on emergency department length of stay and patient disposition, Acad. Emerg. Med. 10 (2003) 127–133.
[5] M.J. Schull, J. Szalai, B. Schwartz, D.A. Redelmeier, Emergency department overcrowding following systematic hospital restructuring: trends at twenty hospitals of ten years, Acad. Emerg. Med. 8 (2001) 1037–1043.
[6] S. Daly, D.A. Campbell, P.A. Cameron, Short-stay units and observation medicine: a systematic review, Med. J. Aust. 178 (2003) 559–563.
[7] M.W. Cooke, J. Higgins, P. Kidd, Use of emergency observation and assessment wards: a systematic literature review, Emerg. Med. J. (2003) 138–142.
[8] G. Damiani, L. Pinnarelli, L. Sommella, V. Vena, P. Magrini, W. Ricciardi, The short stay unit as new option for hospitals: a review of the scientific literature, Med. Sci. Monit. 17 (2011) SR15-9.
[9] I. Scott, L. Vaughan, D. Bell, Effectiveness of acute medical units in hospitals: a systematic review, Int. J. Qual. Health Care 21 (2009) 397–407.
[10] J.C. Moskop, D.P. Sklar, J.M. Geiderman, R.M. Schears, K.J. Bookman, Emergency department crowding, part 2—barriers to reform and strategies to overcome them, Ann. Emerg. Med. 53 (2009) 612–617.
[11] A. Bagust, M. Place, J.W. Posnett, Dynamics of bed use in accommodating emergency admissions: stochastic simulation model, BMJ 319 (2010) 155–158.
[12] T.J. Coats, S. Michalis, Mathematical modelling of patients flow through an accident and emergency department, Emerg. Med. J. 18 (2001) 190–192.
[13] C. Duguay, F. Chetouane, Modeling and improving emergency department systems using discrete event simulation, Simulation 83 (2007) 311–320.
[14] N.R. Hoot, L.J. LeBlanc, I. Jones, S.R. Levin, C. Zhou, C.S. Gadd, D. Aronsky, Forecasting emergency department crowding: a discrete event simulation, Ann. Emerg. Med. 52 (2008) 116–125.
[15] J.K. Cochran, K.T. Roche, A multi-class queuing network analysis methodology for improving hospital emergency department performance, Comput. Oper. Res. 36 (2009) 1497–1512.
[16] L.V. Green, J. Soares, J.F. Giglio, R.A. Green, Using queueing theory to increase the effectiveness of emergency department provider staffing, Am. Emerg. Med. 13 (2006) 61–68.
[17] L. Mayhew, D. Smith, Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments, Health Care Manag. Sci. 11 (2008) 11–21.
[18] M. Ramakrishnan, D. Sier, P.G. Taylor, A two-time-scale model for hospital patient flow, IMA J. Manag. Math. 16 (2005) 197–215.
[19] R.I. Wilkinson, Theories for toll traffic engineering in the USA, Bell Syst. Tech. J. 35 (1956) 421–514.
[20] R.G. Schehrer, A two moment method for overflow systems with different mean holding times, in: Proc of the 15th Int Teletraffic Congr, 1997.
[21] N. Litvak, M. van Rijsbergen, R.J. Boucherie, M. van Houdenhoven, Managing the overflow of intensive care patients, European J. Oper. Res. 185 (2008) 998–1010.
[22] S. Borst, R.J. Boucherie, O.J. Boxma, ERMR: a generalised equivalent random method for overflow systems with repacking, in: Proc of the 16th Int Teletraffic Congr, 1999.
[23] A.M. de Bruin, R. Bekker, L. van Zanten, G.M. Koole, Dimensioning hospital wards using the Erlang loss model, Ann. Oper. Res. 178 (2010) 23–43.
[24] R.W. Wolff, Stochastic Modeling and the Theory of Queues, Prentice Hall, Englewood Cliffs, USA, 1989.
[25] R.W. Wolff, Poisson arrivals see time averages, Oper. Res. 30 (1982) 223–231.
[26] T.J. Ott, The singel server queue with independent GI/G and M/G input streams, Adv. Appl. Probab. 19 (1987) 266–286.
[27] D.A. Stanford, Waiting and interdeparture times in priority queues with Poisson- and general-arrival streams, Oper. Res. 45 (1997) 725–735.